TRANSCRIPTION FACTOR CENTRIC DISCOVERY OF REGULATORY ELEMENTS IN MAMMALIAN GENOMES USING ALIGNMENT-INDEPENDENT CONSERVATION MAPS

Banerjee, Nilanjana[1,2]; Califano, Andrea[1,2,3]
[1]Dept. of Biomedical Informatics, Columbia University, New York, NY;
[2] Joint Centers for System Biology, Columbia University, New York, NY;
[3] Institute of Cancer Genetics, Columbia University, New York, NY
Keywords*:* Network reverse engineering, comparative genomics, DNA binding site analysis, pattern discovery, transcriptional regulation, systems biology

The computational identification of DNA binding sites that have high affinity for a specific transcription factor is an important problem that has only been partially addressed in prokaryotes and lower eukaryotes. Given the higher length of regulatory regions and the relative low complexity of DNA binding signature, however, methods to address this problem in higher order eukaryotes are lacking. Thus, the specific interactions between transcription factors (TFs) and their cognate cis-regulatory elements – ultimately responsible for transcriptional network function – remain largely unmapped in a mammalian context.

In this poster, we propose a novel computational framework, which combines cellular network reverse engineering, integrative genomics, and comparative genomic approaches, to address this problem for a set of human transcription factors. Specifically, we propose to study the regulatory regions of putative orthologous targets of a given transcription factor, obtained by reverse engineering methods, in several mammalian genomes. Highly conserved regions are identified by pattern discovery. Finally DNA binding sites are inferred from these regions using a standard Position Weight Matrices discovery algorithm. By framing the identification of the Position Weight Matrix as an optimization problem over the two parameters of the method, we are able to discover known binding sites for several genes and to propose reasonable signatures for genes that have not been previously characterized.

We first test our approach on a set of six TFs with known TFBS, including *MYC*, *E2F1*, *TFDP1*, *IRF7*, *FOSL1* and *NFkB2*. We then apply our method to two TFs, *BCL6* and *HOXD13,* whose TFBS have not yet been fully characterized. We show that for 4 of the 6 previously characterized transcription factors, our approach is successful in discovering PWMs that closely match the previously known profiles. We identify novel PWMs for *BCL6* and *HOXD13* that are highly consistent with previous knowledge about these genes. The method is of a general nature and can by applied to any set of transcription factors.

This study was supported in part by NCI grant R01 CA109755-01(PI:AC).

E-mail: califano@c2b2.columbia.edu